Analyzing Gambling Addictions:
A Spanish Corpus for Understanding Pathological Behavior

Manuel Couto-Pintos, Marcos Fernández-Pichel,
Mario Ezra Aragón, and David E. Losada
Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS),
Universidade de Santiago de Compostela (USC), Spain
{manuel.couto.pintos,marcosfernandez.pichel,
ezra.aragon,david.losada}@usc.es

Abstract

This work fosters research on the interaction between natural language use and gambling disorders. We have built a new Spanish corpus for screening standardized gambling symptoms. We employ search methods to find on-topic sentences, top-k pooling to form the assessment pools of sentences, and thorough annotation guidelines. The labeling task is challenging, given the need to identify topic relevance and explicit evidence about the symptoms. Additionally, we explore using state-of-the-art LLMs for annotation and compare different sentence search models.

1 Introduction

Compulsive gambling, or gambling disorder, consists of a powerful urge to gamble repeatedly, even when it causes significant financial, emotional, and personal harm. Gambling involves risking something valuable for a potential gain (Humphreys, 2019). For those struggling with this disorder, it can manifest as chasing losses with ever-increasing bets, depleting savings, and accumulating debt. In recognition of the growing global problem of gambling disorders, the World Health Organization (WHO) incorporated them into the International Classification of Diseases (ICD-11) in 2019.[1] Similarly, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), published in 2013, included pathological gambling as a condition warranting further investigation (APA, 2022). Despite the severity of this disorder, many individuals remain undiagnosed or receive treatment late. Existing preventive tools often fall short, thereby opening a critical need for new instruments that can effectively distinguish the spectrum of gaming behaviors, from regular and healthy engagement to hazardous gaming and full-blown disorder (Billieux et al., 2019).

The vast amount of content shared on the Internet and social media presents a unique opportunity for large-scale analysis of psychological traits associated with various disorders. For example, individuals struggling with gambling disorders often find online communities or forums where they connect with others, share their experiences, and seek support from professionals. This data offers a window into the minds and emotions expressed through written communication. Our work analyzes online interactions to identify different publications (posts) containing gambling-related behaviors. We contribute with a new Spanish dataset[2] designed for symptom-level gambling screening. Unlike other disorders, gambling has not garnered enough attention from researchers in computational mental health. The dataset contains sentences labeled according to their relevance to multiple gambling symptoms (Section 3.2). Identifying symptoms at the sentence level helps to detect gambling problems more accurately. It allows models to understand subtle signs, like denial, guilt, or impulsiveness, which might be missed when looking at a whole document. These fine-grained labels potentially produce better monitoring tools and more informative output to end-users (e.g., public institutions screening for signs of gambling).

Candidate sentences were obtained by leveraging 18 ranking methods and applying top-k pooling on their relevance rankings. Four assessors, including a trained Psychologist, annotated each candidate sentence, considering its relevance to each gambling symptom and determining if the sentence offered explicit details about someone (e.g., the post's writer) exhibiting the specific symptom. Additionally, we explore the ability of recent state-of-the-art large language models (LLMs) to annotate the dataset. LLMs can efficiently process massive

[1] https://icd.who.int/en

[2] The dataset and code are available at: https://github.com/citiususc/ludosym

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17610–17619
November 4-9, 2025 ©2025 Association for Computational Linguistics

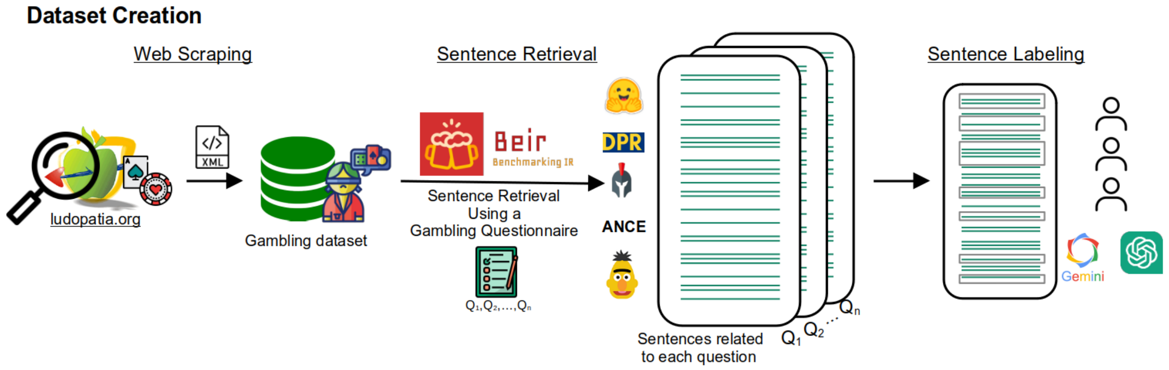


Figure 1: General view of the creation of the gambling sentence corpus. The process includes web scraping, sentence retrieval using BEIR and a gambling questionnaire, and sentence labeling with human assessors and LLMs.

textual datasets, potentially reducing annotation times. We provide insights into their strengths and limitations by comparing LLMs' effectiveness with human assessors. For example, human assessors are susceptible to inherent biases, but LLMs can also inherit biases from their training data. We can summarize our contributions as follows:

1. We construct a dataset in Spanish designed to identify gambling-related evidence. This collection contains sentences tagged at the symptom level. Furthermore, we propose a new search model incorporating domain-related information and yielding competitive performance.

2. We evaluate the ability of LLMs to handle a complex annotation task that involves recognizing symptoms of gambling addiction. This goes beyond standard annotation efforts, requiring the models to understand subtle indicators and contextual cues associated with problematic gambling behavior.

2 Related Work

While Psychology has traditionally dominated the area of gambling disorder research, machine learning methods powered by user-generated data have emerged as complementary tools. A pioneering study by Braverman and Shaffer (2010) employed k-means clustering to identify customers with similar first-month online betting behaviors. The data came from an Internet betting service. Other studies (Dragicevic et al., 2011) have also leveraged data from online gambling platforms, including lottery providers and casinos.

In the literature, multiple aspects have been analyzed, including player behavior (Nicola et al., 2013; Peres et al., 2021), transactions (Ladouceur et al., 2017), and even communications between customers and online gambling support personnel (Haefeli et al., 2011). Notably, the CLEF eRisk workshop marked a significant advancement by leveraging social media data for pathological gambling analysis (Crestani et al., 2022; Ríssola et al., 2021; Parapar et al., 2021, 2022, 2023). Inspired by this line of research, our research builds on the idea that studying online interactions can help us find social media posts where people discuss or engage in gambling behaviors. Creating new datasets in languages other than English will help to investigate new linguistic markers indicative of gambling addictions and advance early detection and intervention strategies.

3 Corpus Creation

The construction of our gambling dataset is oriented explicitly to identifying sentences indicative of gambling-related symptoms. To that end, we leverage a well-established clinical questionnaire (APA, 2022) as our reference for identifying core symptoms within user-generated texts (see Section 3.2). Figure 1 depicts the whole process, consisting of three steps: i) data collection, ii) sentence retrieval, and iii) sentence labeling.

3.1 Data collection

The data originates from ludopatia.org,[3] a forum dedicated to gambling addiction discussions. Several other sources were considered, but this particular source was chosen due to the large amount of available data. Other gambling-related platforms contain user interactions written in Spanish, but do not have a broad user base, or the publications are

[3] https://www.ludopatia.org/forum/default.asp

too short. We segmented the user posts into individual sentences and produced a TREC-formatted collection of 640k sentences contributed by 3,910 unique users, primarily in Spain.[4] Importantly, all extracted sentences are public and crawlable.

3.2 DSM-V Questionnaire

We used the Spanish version of the DSM-V Pathological Gambling Diagnostic Form, which defines nine core symptoms associated to a gambling disorder. Together, these criteria provide a structured taxonomy that captures the cognitive, behavioral, emotional, and social dimensions of gambling disorder. Below, we briefly describe each criterion:

1. Preoccupation with Gambling: Persistent and intrusive thoughts about gambling, including reliving past experiences, planning future episodes, or strategizing ways to obtain money for gambling. This reflects the centrality of gambling in an individual's cognitive life. Related question: Have you often found yourself thinking about gambling (e.g., reliving past gambling experiences, planning the next time you will play, or thinking of ways to get money to gamble)?

2. Increased Betting Amounts: A progressive need to wager larger sums to achieve the same level of excitement, mirroring the tolerance phenomenon observed in substance addictions. Related question: Have you needed to gamble with more and more money to get the excitement you are looking for?

3. Restlessness or Irritability: Emotional distress such as agitation or irritability when attempting to cut back or stop gambling, resembling withdrawal symptoms in substance use disorders. Related question: Have you become restless or irritable when trying to cut down or stop gambling?

4. Gambling to Escape Distress: Using gambling as a maladaptive coping strategy to alleviate negative mood states such as anxiety, depression, or low self-esteem. Related question: Have you gambled to escape from problems or when you are feeling depressed, anxious, or bad about yourself?

5. Chasing Losses: The belief that continuing to gamble after losses will allow recovery of money, often leading to a cycle of escalating bets and debt. Related question: After losing money gambling, have you returned another day to get even?

6. Lying to Conceal Gambling: Dishonesty toward family members, friends, or others to hide the frequency, intensity, or financial consequences of gambling. Related question: Have you lied to your family or others to hide the extent of your gambling?

7. Unsuccessful Control Attempts: Repeated failures to stop, reduce, or regulate gambling behavior despite intentions or attempts. Related question: Have you made repeated unsuccessful attempts to control, cut back, or stop gambling?

8. Jeopardizing Relationships or Opportunities: Gambling that results in the loss or risk of losing important relationships, jobs, or career/educational opportunities. Related question: Have you risked or lost a significant relationship, job, educational, or career opportunity because of gambling?

9. Relying on others for Financial Relief: Seeking financial bailouts from family or friends to cope with gambling-related crises, reflecting both external consequences and interpersonal strain. Related question: Have you sought help from others to provide the money to relieve a desperate financial situation caused by gambling?

3.3 Sentence Retrieval

Our sentence retrieval process relies on the robust and heterogeneous benchmarking-IR (BEIR) framework. This retrieval evaluation benchmark has proven effective for diverse search applications across multiple domains. By rigorously supporting state-of-the-art retrieval models, including linguistic, sparse, dense, late-interaction, and re-ranking architectures, BEIR ensures the reliability and reproducibility of our search experiments.

The sentence retrieval step involves finding candidate sentences indicative of a gambling symptom. Queries are directly formed from the Spanish version of the DSM-V Pathological Gambling Diagnostic Form (e.g., "Have you become restless or irritable when trying to cut down or stop gambling?"). We leveraged 18 search models, including dense and sparse solutions. Dense retrieval encode each query (gambling symptom) and the sentences into a dense embedding space and then identify the sentences closest to the query's vector representation in the embedded space. On the other hand, sparse retrieval or sparse neural models represent texts using sparse vectors (Nguyen et al., 2023). Appendix A.2 presents further details about the search models. Given the rankings of estimated relevant sentences, the top results were merged to form a set of candidates manually assessed for rel-

[4] The sentences contain 33.6 words on average

evance (top-k pooling, with k=100). On average, each pool consisted of 111 candidate sentences per symptom.

3.4 Sentence Labeling

To ensure solid annotations, we recruited four human assessors, including a trained Psychologist. Sentences were deemed relevant only if they directly address the individual's state related to a gambling symptom (or mention someone suffering from the symptom). An on-topic sentence that merely defines a gambling symptom or gives some advice is deemed as non-relevant. A relevant sentence should explicitly indicate that a specific individual is experiencing a gambling symptom. For example, 'Gambling can be risky' would be non-relevant, whereas 'Gambling ruined my life' would be considered relevant.[5] Table 3 (Appendix A.4) shows some examples. We developed a comprehensive set of instructions outlining the assessment criteria (see Figure 4 in Appendix section). These guidelines were provided to the human annotators and subsequently adapted to prompt the LLMs, thus obtaining additional automated judgments. After human annotation, each symptom had an average of 17 relevant sentences, leading to high-quality annotated data. After labeling, the Kappa inter-rater score was 0.39-0.60 (median of 0.43), and Krippendorff's coefficient was 0.44. These values represent a moderate agreement (McHugh, 2012), illustrating the difficulty of this task. This level of agreement aligns with previous studies that manually annotated the presence of mental health symptoms (Pérez et al., 2023; Romero et al., 2024). We consolidated these judgments into a query-relevance file (qrels).[6]

3.5 Automatic Sentence Labeling

We used GPT-4 and Gemini (OpenAI, 2023; Anil et al., 2023), two of the latest and most capable generative AIs, to understand LLMs' capacity to annotate complex psychological markers in text. We prompted the models with the same instructions provided to the human annotators. GPT-4 yielded a Kappa score of 0.25 against the human qrels (see confusion matrix in Appendix A.3). GPT-4 was fairly accurate at identifying irrelevant sentences but produced many false positives (examples in Appendix A.5). On the other hand, Gemini got a Kappa of 0.16 and had a strong tendency to overestimate the number of relevant sentences. This finding is consistent with previous studies that tried to automatically identify other mental health markers in texts (Pérez et al., 2023). These results suggest that LLMs could be exploited to nominate sentences that human assessors then review. For example, if we only present the humans with the sentences labeled by GPT-4 as positive, we would reduce the human workload by approximately 57%, eliminating the need to annotate around 575 sentences. Considering that the average effort per human assessor was 27 hours, this reduction would save around 15 hours of human work. In practice, these savings could be spent on producing deeper assessments.

4 Retrieval of Gambling Symptoms

In Table 1, we present the effectiveness of several baseline search methods (see Appendix A.2) under zero-shot experiments. The table reports two high-precision metrics (P@5 and NDCG@10) and a recall-oriented measure (AP@100). The results illustrate the inherent difficulty of the search task, with most methods achieving modest performance. Among the baselines, the sentence alignment model ST Multi is the best performer for NDCG@10 and AP@100, while the Spanish variant (ST BETO) is the most robust model, detecting the top five relevant sentences. The sentence transformer models (STs) benefit from their training based on sentence corpora (including examples in the target language). In contrast, other models, such as DPR and SPARTA, seem limited by their training based on passages. Although trained with passages, the ANCE model is not far from the best ST models. We hypothesize that its training with hard negatives allows ANCE to capture complex semantic patterns that help generalize to other search tasks. Traditional BM25 lexical retrieval and BM25 combined with re-ranking (BM25+CE) did not perform well, possibly due to vocabulary mismatch. Similarly, expansion models (docT5query) did not perform well for this task.

We also evaluated the inclusion of domain knowledge into the models (Aragon et al., 2023). We re-ranked the ST Multi top-k results with a custom cross-encoder that uses a domain-adapted language model named LudoBETO (Appendix A.1).

[5] Since we employ pooling to retrieve candidate sentences, the same sentence may end up in multiple pools and, thus, it might be deemed relevant for several symptoms.

[6] A sentence is relevant when at least two annotators marked it as so

Models	NDCG@10	AP@100	P@5
BM25	.122	.037	.133
SPARTA	.121	.030	.133
docT5query	.037	.011	.044
DPR	.018	.004	.022
ANCE	.234	.071	.244
ST BETO	.253	.086	.311
ST Multi	.306	.117	.266
ST Multi-E5-base	.116	.063	.133
ST Multi-E5-large	.126	.058	.089
ST Multi-E5-small	.209	.086	.178
BM25+CE	.146	.045	.133
ST Multi+ +CE ludoBETO	**.376**	**.165**	**.378**

Table 1: Search Performance Results

The cross-encoder was trained using contrastive learning with in-domain sentences. This strategy (ST Multi+CE ludoBETO) yielded the best results overall, suggesting that including information related to the domain helps the model during the search task. Other ablation variants were also tested (with no domain adaptation and with no initial candidate generation by ST Multi), but they were inferior to ST Multi+CE ludoBETO.

4.1 Occurrence of Gambling Symptoms

Figure 2 (left) reports the proportion of relevant sentences for each DSM-V criterion. Topics related to money or help, like "*Increased Betting Amounts*" or "*Relying on Others for Financial Relief*" are scarce. Instead, other topics, such as "*Preoccupation with Gambling*" or "*Lying to Conceal Gambling*", are more prevalent. This is an expected outcome since the notion of relevance is intricate, and many candidate sentences do not reveal clear evidence. The right part of Figure 2 shows the P@5 results with ST BETO. Some frequent symptoms (e.g., "*Preoccupation with Gambling*") exhibit low performance. In contrast, other less frequent symptoms (e.g., "*Restlessness or Irritability*") seem easier to retrieve. This opens up an interesting line of research that is oriented toward adapting the search methods to the characteristics of specific symptoms.

4.2 Error Analysis

In this section we provide some examples that illustrate how hard it is for the model to retrieve relevant sentences for some of the topics of the questionnaire. For example, for the symptom "*Jeopardizing Relationships or Opportunities*", the model retrieves on-topic sentences such as 'Loss of significant work relationships or opportunities in their studies or career due to gambling' or 'Significant interpersonal relationships, work or career opportunities have been jeopardized or lost because of gambling'. These sentences merely define this gambling criterion, and one cannot infer from these words that a specific individual is experiencing that situation.

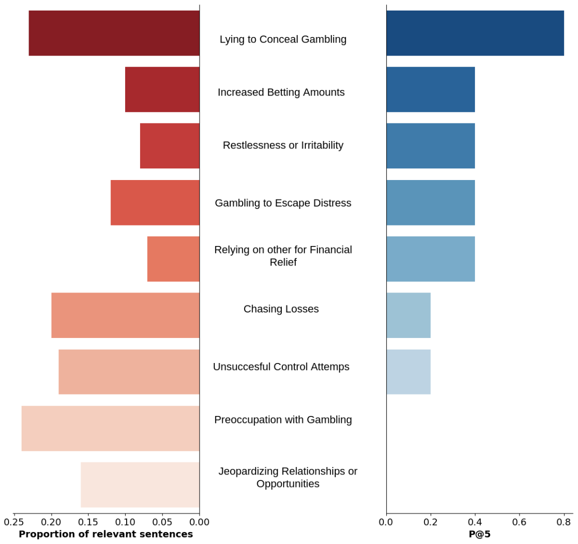


Figure 2: Proportion of relevant sentences (left) and P@5 performance (right) for ST BETO model and each DSM-V criterion.

Similarly, for the symptom "*Relying on Others for Financial Relief*", some retrieved sentences ('Need to borrow money to survive due to losses caused by gambling' or 'I suggest that if you feel the need to do something, you contact the entities that carry out social work on the island and offer to collaborate') do not explicitly reflect that a specific individual is experiencing that situation.

5 Conclusion and Future Work

This study has introduced a new language resource for fostering research on standardized gambling symptoms. The dataset includes sentences tagged based on their relevance to crucial gambling assessment criteria. We used advanced search techniques to locate relevant sentences, top-k pooling to create assessment sentence pools and annotation guidelines. Furthermore, we showed that LLMs could act as filters to remove non-relevant sentences from the candidate pools, but they still struggle to infer psychological states from textual cues. In future work, we aim to continue expanding this corpus, exploring the potential of other models, and implementing hybrid annotation approaches combining the strengths of LLMs and humans.

Limitations

Our main goal is to enhance societal benefits by advancing the understanding of pathological gambling. This research does not aim to diagnose individual mental health conditions. It is crucial to acknowledge that the dataset may exhibit biases and limitations typical of social media studies. For instance, the dataset does not reflect the entire global population, focusing on people from different Spanish-speaking countries. Besides this geographical orientation, the dataset may have biases related to population segments. For example, elderly individuals who do not use online platforms or users with private accounts cannot be included in this sample.

We also recognize that the search methods tested here are preliminary, and our retrieval experiments should be interpreted as initial tests to provide a reference set of baseline models. Optimizing retrieval effectiveness was not our goal. For example, generating effective queries from DSM-V questions was out of the scope of our work, but it represents a promising avenue for future development.

While the sample size is limited, we view this dataset as a starting point and plan to expand it further. For example, we presented here preliminary results exploring the potential of LLMs as automatic annotators of gambling symptoms. As our results suggest, these models show promise in helping reduce the number of non-relevant sentences, allowing human annotators to focus on a more manageable set of relevant sentences. We plan to continue improving and scaling this dataset in future iterations.

Additionally, this study exclusively considered publications in Spanish. Nevertheless, our strategies are highly adaptable and can be seamlessly applied across various languages, facilitating the exploration of how gambling symptoms manifest in diverse countries and cultures. This cross-linguistic flexibility could enable a deeper understanding of global cultural and societal factors influencing gambling behaviors. With this in mind, in future research, we plan to broaden our scope to include other languages and social media platforms and to create effective strategies for understanding and reducing biases. Researchers and practitioners using our data should recognize these possible biases and take steps to ensure fairness.

Challenges

This task is challenging; symptoms are sometimes expressed indirectly or subtly. Phrases like 'just one more bet' or 'I can quit anytime' might indicate denial or compulsive behavior, but can also be interpreted casually. Models struggle to differentiate between benign mentions and symptomatic behavior. This leads to false positives where neutral statements are flagged as problematic. Related to this, some symptoms are directly associated with cue words (e.g., 'bet more' for the 'Increased Betting Amounts' symptom), while others are more subtle. This leads to notable differences in symptom detection effectiveness.

Identifying gambling-related symptoms in Spanish presents unique challenges that dense retrieval models may struggle with, including ambiguity and polysemy. The Spanish language often uses words with multiple meanings depending on context. For example, 'Apuesto todo el tiempo' can mean 'I bet all the time' (a clear gambling symptom) but could also imply confidence or conviction about a situation that is not related to gambling. Dense retrieval models trained on English data (or multilingual models trained with low proportions of Spanish examples) might not handle such polysemy well, leading to false positives or missed detections. Furthermore, Spanish-speaking regions have diverse idiomatic expressions that vary largely by country; models without exposure to regional idioms may struggle to recognize symptoms accurately.

Ethic Statement

The primary goal of this resource is to advance technologies that detect symptoms associated with gambling behaviors. Neuman et al. (2012) emphasized that these innovative methods and tools should be considered enhancements to professional judgment, not replacements. Automated screening methods should be viewed as digital aids that reduce the strain on public health systems.

The dataset is available upon request, and we adhered to strict ethical guidelines, particularly concerning AI's ethical development. We anonymized the users' accounts to safeguard privacy. Specific textual excerpts were de-identified to ensure that publications cannot be traced back to individual users. This process involves removing or altering identifiers that could potentially reveal the identity of the users. By anonymizing data, we aim to reduce risks and protect user identities while still enabling valuable insights to be derived. To prevent misuse, this collection can only be used for research purposes. Anyone interested in accessing this data must fill out a user agreement and send it to the authors. Furthermore, we are committed to continuously reviewing and enhancing our data anonymization and protection practices. We stay informed about best practices and technological advancements to ensure our methods are up-to-date and effective.

This study did not involve interacting with social media users, such as providing health advice. Instead, it was an observational study utilizing publicly available data, respecting the site's access and crawling rules. Since we only used public data and did not contact social media users, this research was considered exempt from our IRB review.

Acknowledgments

The authors thank the support obtained from MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF) and Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF).

The third author also thanks the support obtained from the Juan de la Cierva Grant (JDC2023-052296-I), funded by MCIN/AEI/10.13039/501100011033 and by the FSE+.

References

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

American Psychiatric Association APA. 2022. Diagnostic and statistical manual of mental disorders: Dsm-5. *American psychiatric association*.

Mario Aragon, Adrián Pastor López Monroy, Luis Gonzalez, David E Losada, and Manuel Montes. 2023. Disorbert: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318.

Joël Billieux, Maèva Flayelle, Hans-Juergen Rumpf, and Dan J Stein. 2019. High involvement versus pathological involvement in video games: a crucial distinction for ensuring the validity and utility of gaming disorder. *Current Addiction Reports*, 6:323–330.

Julia Braverman and Howard J Shaffer. 2010. How do gamblers start gambling: identifying behavioural markers for high-risk internet gambling. *Eur J Public Health*, 22(2):273–278.

Fabio Crestani, David E Losada, and Javier Parapar. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, volume 1018. Springer Nature.

Simo Dragicevic, George Tsogas, and Aleksandar Kudic. 2011. Analysis of casino online gambling data in relation to behavioral risk markers for high-risk gambling and player protection. *International Gambling Studies*, 11(3):377–391.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Joerg Haefeli, Suzanne Lischer, and Juerg Schwarz. 2011. Early detection items and responsible gambling features for online gambling. *International Gambling Studies*, 11(3):273–288.

Gary Humphreys. 2019. Sharpening the focus on gaming disorder. *World Health Organization. Bulletin of the World Health Organization*, 97:382–383.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Robert Ladouceur, Paige Shaffer, Alex Blaszczynski, and Howard J. Shaffer. 2017. Responsible gambling: a synthesis of the empirical evidence. *Addiction Research & Theory*, 25(3):225–235.

Xiong Lee, Xiong Chenyan, Li Ye, Tang Kwok-Fung, Liu Jialin, Bennett Paul, Ahmed Junaid, and Overwijk Arnold. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Preprint*, arXiv:2007.00808.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Yair Neuman, Yohai Cohen, Dan Assaf, and Gabbi Kedma. 2012. Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine*, 56(1):19–25.

Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. A unified framework for learned sparse retrieval. In *Advances in Information Retrieval*, pages 101–116, Cham. Springer Nature Switzerland.

Adami Nicola, Benini Sergio, Boschetti Alberto, Canini Luca, Maione Florinda, and Temporin Matteo. 2013. Markers of unsustainable gambling for early detection of at-risk online gamblers. *International Gambling Studies*, 13(2):188–204.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docttttquery. *Online preprint*, 6(2).

OpenAI. 2023. Gpt-4 technical report. *arXiv:submit/4812508*.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2021. Overview of erisk 2021: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 324–344, Cham. Springer International Publishing.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2022. Overview of erisk 2022: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 233–256, Cham. Springer International Publishing.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 294–315, Cham. Springer Nature Switzerland.

Fernando Peres, Enrico Fallacara, Luca Manzoni, Mauro Castelli, Aleš Popovič, Miguel Rodrigues, and Pedro Estevens. 2021. Time series clustering of online gambling activities for addicted users' detection. *Applied Sciences*, 11(5).

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar, and David E. Losada. 2023. Depresym: A depression symptom annotated corpus and the role of llms as assessors of psychological markers. *Preprint*, arXiv:2308.10758.

Esteban A. Ríssola, David E. Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2).

Alba M Mármol Romero, Adrián Moreno Muñoz, Flor Miriam Plaza Del Arco, M Dolores Molina-González, María Teresa Martín Valdivia, L Alfonso Urena Lopez, and Arturo Montejo Ráez. 2024. Mentalriskes: A new corpus for early detection of mental disorders in spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11204–11214.

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics.

A Appendix

A.1 ludoBETO

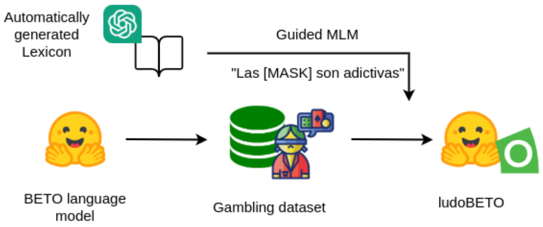


Figure 3: General diagram of the domain adaptation process. It consists of a guided masked language model supported by a domain lexicon generated by GPT-4.

We domain-adapted the BETO language model to the pathological gambling domain.[7] As seen in Figure 3, we asked GPT-4 to generate a pathological gambling lexicon and then adapted the BETO model for masked language modeling using sentences from the gambling dataset. During the training process, we prioritized masking words from the lexicon, supplementing the mask with random words as needed to reach a total of 15% masked words per sentence. The model was trained during four epochs with a learning rate of 2×10^{-5}.

We tested ludoBETO's language modeling by masking and predicting random words on a separate test set. In these experiments, ludoBETO had a perplexity of 9.02, while BETO's perplexity was 33.51. This result indicates that our model is acquiring domain knowledge without losing general language modeling capabilities.

A.2 Retrieval Models

The experiments were run with a GPU GeForce RTX 3070. We implemented the models using the BEIR framework and SentenceTransformer, version 2.7.0. In particular, given each query built from the DSM-V form, we searched for the top-k (k=100) sentences using the following models:

- BM25: popular probabilistic ranking function for document retrieval that uses term frequency, term saturation, and document length normalization to score documents.
- docT5query (Nogueira et al., 2019): document expansion technique that leverages a sequence-to-sequence T5 model (fine-tuned

[7] https://huggingface.co/citiusLTL/ludoBETO

on MS MARCO) to generate synthetic queries and append them to the original documents. The expanded Liew Zi Xianindex supports a traditional lexical search.

- SPARTA (Zhao et al., 2021): SPARTA learns a sparse representation that can be efficiently implemented as an inverted index. The resulting representation enables scalable neural retrieval that does not require expensive approximate vector search.
- DPR (Karpukhin et al., 2020): Dense passage retrieval approach where a dual-encoder learns the association between queries and documents.
- ANCE (Lee et al., 2020): Approximate Nearest Neighbor Negative Contrastive Estimation. Constructs negative examples from an Approximate Nearest Neighbor corpus index.
- ST BETO: Sentence transformers model based on a BERT variant trained on a big Spanish corpus.
- ST Multi: Sentence transformers model based on the multilingual-MiniLM-L12-v2 model.
- ST Multi-E5-base: Sentence transformers model based on the Multi-E5-base model.
- ST Multi-E5-large: Sentence transformers model based on the Multi-E5-large model.
- ST Multi-E5-small: The sentence transformers model is used on the Multi-E5-small model.
- BM25+CE: uses BM25 for initial candidate retrieval and, next, a cross-encoder for re-ranking the top-k ($k = 10$) examples. We used *cross-encoder/ms-marco-MiniLM-L-6-v2* as a cross-encoder.
- ST Multi+CE ludoBETO: combines ST Multi for initial candidate retrieval and a custom cross-encoder for re-ranking the top-k ($k = 10$) items. The cross-encoder was trained using simple contrastive learning with in-domain sentences (Gao et al., 2021).

Contrastive learning teaches models to distinguish between thematically related sentences that have highly different meanings in the context of gambling screening. For example, "Gambling ruined my life" (Gambling symptom) vs "Gambling can be risky" (Neutral). Models can thus handle ambiguity and context better by learning to map similar sentences closer together in vector space while pushing dissimilar ones apart.

A.3 Automatic Labeling

Table 2 shows the confusion matrix of GPT-4 against the human-generated qrels.

	Pred. non-rel	**Pred. rel**
Human non-rel	465	379
Human rel	10	145

Table 2: Confusion Matrix (GPT-4 vs. humans)

A.4 Example of Labeled Sentences

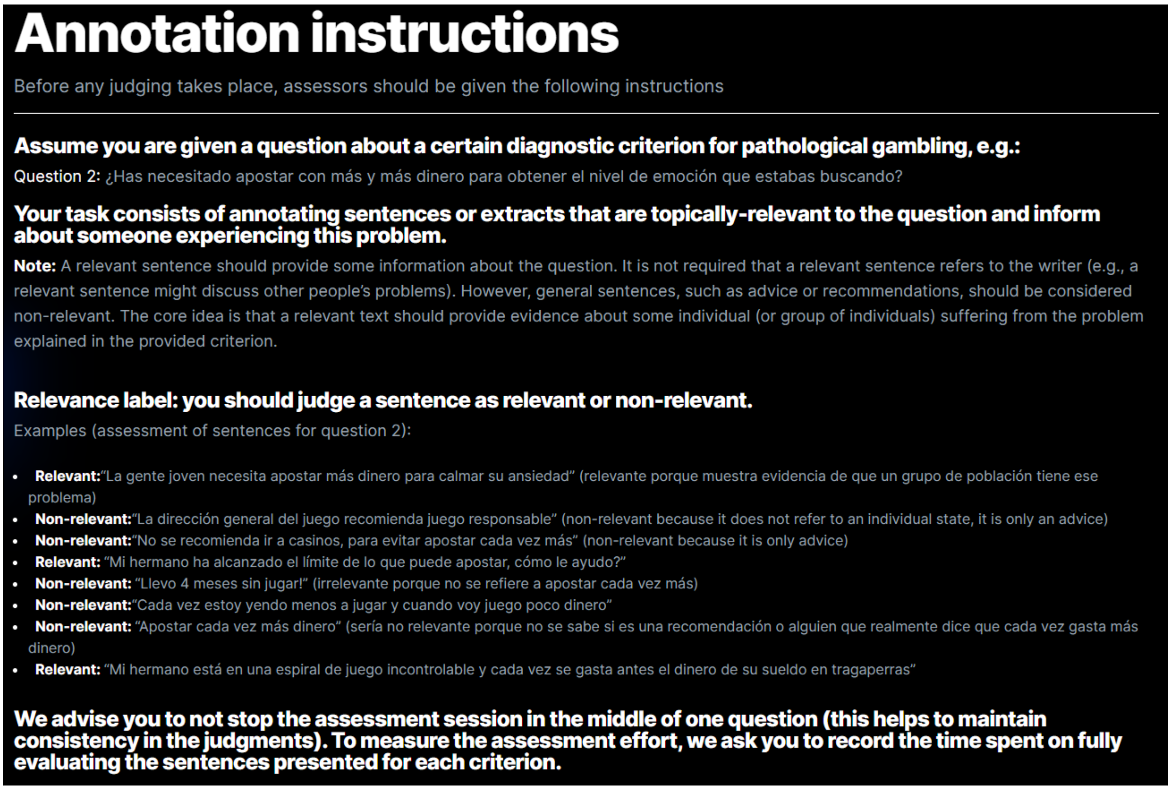
We selected four human assessors with different backgrounds: a trained Psychologist, a PhD student, and two postdocs (the PhD student and the two postdocs have a background in Computer Science and Mental Health analysis). Table 3 shows examples of relevant and non-relevant sentences in Spanish (and their English translations). The first sentence is considered non-relevant because it is on-topic. Still, it just defines one gambling criterion (and one cannot infer from these words that a specific individual is experiencing that situation).

Relevance	**Sentence**
0	"*Necesidad de apostar sumas crecientes para sentir excitación*" **translation:** "*Need to gamble increasing sums to feel excitement*"
1	"*Perdía el raciocinio apostando cantidades cada vez mayores para sentir estímulos más intensos...*" **translation:** "*I lost my reasoning by gambling increasing amounts to feel more intense stimuli...*"

Table 3: Examples of sentences for the symptom *Increased Betting Amounts*. Sentences are paraphrased for anonymity purposes.

A.5 LLMs qualitative analysis

Table 4 illustrates some examples of false positives produced by the LLMs for the symptom "*Unsuccessful Control Attempts*". For instance, GPT-4 labels one question and one piece of advice as relevant sentences. However, these two texts do not provide evidence that someone is suffering from the target symptom. On the other hand, Gemini's



Annotation instructions

Before any judging takes place, assessors should be given the following instructions

Assume you are given a question about a certain diagnostic criterion for pathological gambling, e.g.:

Question 2: ¿Has necesitado apostar con más y más dinero para obtener el nivel de emoción que estabas buscando?

Your task consists of annotating sentences or extracts that are topically-relevant to the question and inform about someone experiencing this problem.

Note: A relevant sentence should provide some information about the question. It is not required that a relevant sentence refers to the writer (e.g., a relevant sentence might discuss other people's problems). However, general sentences, such as advice or recommendations, should be considered non-relevant. The core idea is that a relevant text should provide evidence about some individual (or group of individuals) suffering from the problem explained in the provided criterion.

Relevance label: you should judge a sentence as relevant or non-relevant.

Examples (assessment of sentences for question 2):

- **Relevant:**"La gente joven necesita apostar más dinero para calmar su ansiedad" (relevante porque muestra evidencia de que un grupo de población tiene ese problema)
- **Non-relevant:**"La dirección general del juego recomienda juego responsable" (non-relevant because it does not refer to an individual state, it is only an advice)
- **Non-relevant:**"No se recomienda ir a casinos, para evitar apostar cada vez más" (non-relevant because it is only advice)
- **Relevant:** "Mi hermano ha alcanzado el límite de lo que puede apostar, cómo le ayudo?"
- **Non-relevant:** "Llevo 4 meses sin jugar!" (irrelevante porque no se refiere a apostar cada vez más)
- **Non-relevant:**"Cada vez estoy yendo menos a jugar y cuando voy juego poco dinero"
- **Non-relevant:** "Apostar cada vez más dinero" (sería no relevante porque no se sabe si es una recomendación o alguien que realmente dice que cada vez gasta más dinero)
- **Relevant:** "Mi hermano está en una espiral de juego incontrolable y cada vez se gasta antes el dinero de su sueldo en tragaperras"

We advise you to not stop the assessment session in the middle of one question (this helps to maintain consistency in the judgments). To measure the assessment effort, we ask you to record the time spent on fully evaluating the sentences presented for each criterion.

Figure 4: Annotation instructions that assessors read before submitting their judgment. Complete instructions are available at https://gambling-guidelines.vercel.app/

example shows that the LLM did not distinguish between successful and unsuccessful attempts (the sentence is on-topic but describes a successful control attempt rather than an unsuccessful attempt). These examples illustrate LLMs' difficulty in correctly interpreting this fine-grained notion of relevance.

Model	Sentence
GPT-4	"*¿Alguna vez ha tratado de disminuir el tiempo que dedica al juego?*" **translation:** "*Have you ever tried to reduce the time you spend gambling?*" "*Lo ideal sería poder controlarse al jugar y conocer tus límites, aunque esto resulta casi imposible.*" **translation:** "*Ideally, you should be able to control yourself when playing and know your limits, although this is almost impossible.*"
Gemini	"*He logrado dejar de jugar y soy consciente de que no debo hacerlo nunca más. Pasan meses y sigo sin jugar.*" **translation:** "*I have managed to stop playing and know I should never do it again. Months go by and I still don't play*"

Table 4: Examples of false positives for the symptom "*Unsuccessful Control Attempts*".

A.6 Annotation Guidelines

Figure 4 shows the annotation instructions given to the assessors.